DOCUMENT RESUME

ED 177 193 TM 009 685

AUTHOR Squire, Lisa S.; Sattler, Jerome M.
TITLE Examiner Scoring of Three Subtests of the McCarthy Scales of Children's Abilities.
PUB DATE Apr 79
NOTE 12p.; Paper presented at the Annual Meeting of the Western Psychological Convention (San Diego, California, April, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Early Childhood Education; Freehand Drawing; *Intelligence Tests; Preschool Tests; *Scoring; *Testing Problems; *Test Reliability; *Visual Measures; Vocabulary Development
IDENTIFIERS *Interrater Reliability; *McCarthy Scales of Childrens Abilities

ABSTRACT

Two studies investigated examiner scoring agreement on the Oral Vocabulary, Draw-A-Design, and Draw-A-Child subtests of the McCarthy Scales of Children's Abilities. In study one, ten examiners scored four randomly selected normal children's protocols. Unanimous agreement occurred on scores for 52% of the Oral Vocabulary responses, 39% of the Draw-A-Design drawings, and 72% of the Draw-A-Child body parts. In study two, 11 examiners scored 50 pre-selected ambiguous Oral Vocabulary responses. Considerable scoring disagreement occurred, with unanimous agreement found for only 14% of the responses. The results indicated that ambiguous vocabulary responses were as difficult to score on the McCarthy Scales as they are on other individually administered intelligence tests. The Draw-A-Design subtest was also difficult to score. However, Oral Vocabulary responses and Draw-A-Child drawings obtained in a routine administration from normal children had relatively few scoring problems. It was concluded that considerable judgment is required in scoring some McCarthy subtests. (Author/MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

-ED177193

Examiner Scoring of Three Subtests of the McCarthy Scales of Children's Abilities

Lisa S. Squire

University of North Dakota

Jerome M. Sattler

San Diego State University

(Paper presented at Western Psychological Convention, San Diego, Cal., April, 1979)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Lisa S. Squire

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An inspection of the 18 subtests in the McCarthy Scales of Children's Abilities indicated that three subtests may present special scoring difficulties. These include the Oral Vocabulary part of the Word Knowledge subtest, the Draw-a-Design subtest, and the Draw-a-Child subtest. These subtests require considerable examiner judgment in deciding on the appropriate score. Some unreliability on these subtests may be expected because previous studies with other individually administered intelligence tests have demonstrated differences among examiners in their scoring decisions.

We conducted two different, but related, investigations. In the first study, we investigated scoring agreement on subtest responses as they actually were given to the three subtests in a standard test administration. Four different protocols were selected for this purpose. In the second study, we investigated the extent of agreement in scoring ambiguous Oral Vocabulary responses. We selected reponses that appeared to pose special scoring problems.

During the Spring of 1978, 68 children ranging in age from 4 years, 0 months to 8 years, 6 months were individually administered the Oral Vocabulary, Draw-a-Design, and Draw-a-Child subtests of the McCarthy Scales. The children were enrolled in eight preschools and private elementary schools in the San Diego, California area.

TM009 685

In Study I, the 68 protocols were screened, and we selected only those which had 8 or more responses to the Oral Vocabulary words. This criterion was used because we wanted protocols which had responses to most of the subtest items. From the pool of 68 protocols, 16 met the criterion, and from these 16 protocols, 4 were randomly selected. These four protocols came from children who were seven or eight years old. In Study II, we selected five ambiguous responses for each of the ten Oral Vocabulary words from the 68 protocols.

For Study I, we constructed a rating form which consisted of three parts. The first part gave directions for scoring the Oral Vocabulary, Draw-a-Design, and Draw-a-Child responses. Each examiner was furnished with a copy of the McCarthy scoring guidelines and was instructed to study them carefully. Examiners were asked to indicate how confident they were of scores they assigned to each response by using the following scale: A = Very certain of score, B = Somewhat certain of score, C = Minimally certain of score, and D = Guess. The second part of the form consisted of the four different children's protocols, each containing 10 Oral Vocabulary responses, 9 Draw-a-Design drawings, and 1 Draw-a-Child human figure drawing. To control for order effects, the order of presentation of the four protocols was randomized separately for each examiner. However, the three subtests were always presented in the same order as they appear in the McCarthy, namely, Oral Vocabulary, Draw-a-Design, and Draw-a-Child. The third part of the rating form was a personal data sheet that asked about the examiner's sex, education, position, and testing experience.

For Study II, the rating form was similar to that used in Study I. However, in the first part instructions were given only for scoring Oral Vocabulary responses, and in the second part only the 50 ambiguous Oral Vocabulary respon-

ses were presented. The third part of the form was the same.

Twenty-four graduate students in psychology, each of whom had completed a course in individual intelligence testing, agreed to participate in the study. They were randomly assigned to either Study I or II. Of the 12 raters assigned to Study I, 10 (3 male and 7 female) returned the rating forms. None had adminstered the McCarthy Scales. Of the 12 raters assigned to Study II, 11 (2 male and 9 female) returned the rating forms. None had administered the McCarthy Scales.

Our primary analysis for both studies focused on rates of scoring agreement for the subtests. Tallies were made of examiner scores, and cumulative percentages of scoring agreement were tabulated for each subtest.

We also evaluated the reliability of the examiner ratings by using the intraclass correlation procedure. In addition, the relationship between level of scoring agreement and confidence ratings was evaluated by using the Spearman rank order correlation procedure in each study.

In Study I the results showed that rates of scoring agreement were usually highest for Draw-a-Child, followed in decreasing order by Oral Vocabulary and Draw-a-Design (see Table 1). For example, unanimous agreement occurred on scores for 72% of the Draw-a-Child body part reproductions, 52% of the Oral Vocabulary responses, and 39% of the Draw-a-Design drawings. An agreement criterion of 8 out of 10 found the raters agreeing on 92% of the Draw-a-Child body parts, 85% of the Oral Vocabulary responses, and 62% of the Draw-a-Design drawings.

An analysis of rates of scoring agreement for each item in the three subtests indicated that agreement rates were uniformly high for the 10 items in the Draw-a-Child subtest. For Oral Vocabulary, only the last two words present-

ed some difficulty in scoring. However, on the Draw-a-Design subtest, Designs 4 through 9 were particularly difficult to score, with agreement rates ranging from 50% to 85%, while Designs 1 through 3 yielded agreement rates of 100% (see Table 2).

The intraclass correlation procedure indicated that raters achieved satisfactory reliability in scoring the Oral Vocabulary (r = .91, $p < .01$) and Draw-a-Child (r = .81, $p < .01$) subtests, but poor reliability in scoring the Draw-a-Design (r = .22, $p < .05$) subtest. Raters were extremely confident of their scores, as indicated by mean confidence ratings of 3.74 for Draw-a-Child, 3.60 for Oral Vocabulary, and 3.54 for Draw-a-Design on the 4-point scale. Rank order correlations indicated that rates of scoring agreement were significantly associated with confidence ratings, as rhos ranged from .58 to .94 for the four protocols on the three subtests.

In Study II, which focused on scoring the ambiguous Oral Vocabulary responses, considerable scoring disagreement was found (see Table 3). For example, unanimous agreement occurred for only 14% of the responses. An agreement criterion of 8 out of 11 resulted in agreement on scores for 60% of the responses.

The intraclass correlation procedure indicated that moderate interrater reliability (r = .51, $p < .01$) was achieved in scoring the ambiguous Oral Vocabulary responses. Raters were extremely confident of their scores, as indicated by a mean confidence rating of 3.33. A rank order correlation indicated that scoring agreement rates were moderately associated with confidence ratings (rho = .39, $p < .01$).

The results point out that when responses were scored as they were obtained in a standard administation, two of the three McCarthy subtests studied in this

investigation, Draw-a-Child and Oral Vocabulary, were found to produce few scoring problems for examiners. However, considerable examiner unreliability was found in scoring the Draw-a-Design subtest and ambiguous Oral Vocabulary responses.

Now, why did Draw-a-Design produce lower rates of scoring agreement than those reported from studies with the Bender-Gestalt, using the Koppitz scoring system? An inspection of the McCarthy and Koppitz scoring systems revealed that there are almost twice as many scoring criteria for the nine McCarthy designs as there are for the nine Bender designs (47 criteria vs. 25 criteria). The McCarthy scoring guidelines are much more detailed, with some designs requiring the examiner to make as many as nine specific scoring decisions. This difference in the number of scoring criteria might account in part for the differences in interrater reliability.

The more complex designs (i.e., 4 through 9) are particularly difficult to score. One possible reason for this difficulty is that the rater uses two sets of criteria, with six to eight scoring rules, in arriving at scores for Designs 4 through 9, but only one set, with two scoring rules, for Designs 1 through 3. Thus, the greater number of distinctions and judgments required for the more complex designs promotes variability in scores given by raters. The results suggest that is is especially difficult to score designs of older children because of the ambiguities arising over the thorny distinctions between 2- and 3-point responses. Protocols obtained from younger children, who only draw the first three designs, are likely to result in high examiner scoring agreement.

A possible reason for the low rates of agreement in scoring ambiguous Oral Vocabulary responses is that raters must make fine distinctions in determ-

ining whether the scoring guidelines have been fulfilled. The rater must determine whether a response demonstrates that the child has adequate (2 points), partial (1 point), or no knowledge (0 points) of the meaning of the word. The results suggest that raters do not agree on where these fine distinctions should be made in scoring ambiguous Oral Vocabulary responses given the current guidelines.

Two children's responses given to the word shrink illustrate the difficulties involved in scoring ambiguous Oral Vocabulary responses. To what extent does a child understand the meaning of shrink when he responds with "It means shrink things down to size" or with "Something when it's snowing and build a snowman and it shrinks when the sun comes up"? Correct 2-point responses in the manual include "Something gets smaller," "A big thing turns into a little thing," and "Washing makes things littler." Both of the children's responses resulted in considerable scoring disagreement. For the former response, four rater assigned a score of 2, five raters, 1(Q), and two raters, 0(Q). For the latter response, five raters assigned a score of 2, five raters, 1(Q), and one rater, 0. Both responses give a hint of understanding, but the extent to which the child understands the meaning of shrink is not clear in either response.

When a child defines a word using the word in his definition, the examiner, according to the McCarthy manual, must ask the child to define the word in other terms. However, the examiner is placed in a difficult position when the child still uses the word in his definition after being told to redefine the word. The fact that the child is able to describe the word in terms of its usage indicates some knowledge of its meaning, but to what extent? While ambiguous responses require probing by the examiner, problems arise when a child cannot elaborate an ambiguous response or when the response is elabor-

ated but still vague. Yet the examiner must score the response using the available guidelines. This problem, of course, also occurs in scoring vocabulary responses on other individually administered intelligence tests, such as the WISC-R, WPPSI, and Stanford-Binet.

The high interrater reliability coefficients for Draw-a-Child were consistent with coefficients reported in studies of interrater reliability for the Draw-a-Man Test. Of the two visual-motor tests investigated in the present study, interrater reliability for Draw-a-Child was considerably higher than for Draw-a-Design. One possible reason for this difference is that body part may be easier to score than geometric designs. Raters are more able to agree on whether a body part is present (like whether or not a neck is present) than on whether a design has a particular shape (like whether circles are more nearly circular than oval).

The scoring disagreements encountered on Oral Vocabulary and Draw-a-Design do not appear to be a shortcoming of the McCarthy manual. The scoring criteria for both subtests are clear and easy to understand. However, the scoring guidelines cannot cover all possible varieties of responses. Raters will always be required to use judgment.

Overall, the raters were extremely confident of their scores, although they were less confident in scoring responses that led to greater scoring disagreement. Examiners should be made aware of the potential difficulties in scoring the Draw-a-Design and Oral Vocabulary subtests of the McCarthy Scales.

A possible shortcoming of this study is that raters had limited experience in administering individual intelligence tests and visual-motor tests. However, all raters had completed one or more courses in individual assessment and had administered a number of Bender-Gestalt tests. Research on examiner

experience has indicated that graduate student examiners' rates of scoring agreement closely parallel those of more experienced examiners. In addition, the fact that rates of agreement were 100% for the three easiest Draw-a-Design stimuli and 95% to 100% on 8 of 10 of the Draw-a-Child body parts indicates that the raters were extremely careful in following the guidelines in the manual. Therefore, we believe that the results do provide important information about scoring the McCarthy Scales.

Table 1

Cumulative Percent of Scoring Agreement Across Four Protocols for Three McCarthy Subtests

Interrater agreement criterion	Cumulative percent		
	Oral Vocabulary	Draw-A-Design	Draw-A-Child
10 out of 10	52	39	72
9 out of 10	75	47	85
8 out of 10	85	62	92
7 out of 10	98	78	93
6 out of 10	100	78	100
5 out of 10		85	
4 out of 10		96	
3 out of 10		100	

Table 2

Percent of Raters Giving the Same Score to Each Item on Three McCarthy Subtests

Item	Oral Vocabulary	Draw-A-Design	Draw-A-Child
1	98	100	95
2	98	100	100
3	90	100	98
4	98	85	80
5	35	65	100
6	90	83	95
7	88	60	95
8	88	73	100
9	82	50	85
10	82	-	100

Note. Percent of agreement is based on the average agreement for the four protocols.

Table 3

Cumulative Percent of Scoring Agreement for 50 Ambiguous McCarthy Oral Vocabulary Responses

Interrater agreement criterion	Cumulative percent
11 out of 11	14
10 out of 11	28
9 out of 11	42
8 out of 11	60
7 out of 11	76
6 out of 11	88
5 out of 11	92
4 out of 11	100